Thomas Dobereiner

A Data-Driven Approach to Student Performance

Brazil

21/11/2018

Thomas Dobereiner

A Data-Driven Approach to Student Performance

Undergraduate Thesis presented to the Federal University of Santa Catarina as a requisite for the bachelor degree of Electronics Engineering.

Federal University of Santa Catarina - UFSC

Eletronics Engineering

Supervisor: Richard Demo Souza

Co-supervisor: Eduardo Luiz Ortiz Batista

Brazil

21/11/2018

Thomas Dobereiner
A Data-Driven Approach to Student Performance/ Thomas Dobereiner. – Brazil, 21/11/2018-
50 p. : il. (algumas color.) ; 30 cm.

Supervisor: Richard Demo Souza

Undergraduate Thesis – Federal University of Santa Catarina - UFSC
Eletronics Engineering , 21/11/2018.

1. Data Science 2. Education I. Richard Demo Souza II. Federal University of Santa Catarina III. Bachelor of Electronics Engineering IV. Data-Driven Approach to Student Performance

Thomas Dobereiner

A Data-Driven Approach to Student Performance

Esta Trabalho foi julgada adequada para obtenção do Título de Bacharel em Engenharia Eletrônica e aprovada em sua forma final pela Banca Examinadora

Florianópolis, 03 de Dezembro de 2018.

Prof. Jefferson Luiz Brum Marques, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Richard Demo Souza, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Marcio Cherem Schneider, Dr.
Universidade Federal de Santa Catarina

Prof. Walter Pereira Carpes Junior, Dr.
Universidade Federal de Santa Catarina

Thomas Dobereiner

A Data-Driven Approach to Student Performance

Undergraduate Thesis presented to the Federal University of Santa Catarina as a requisite for the bachelor degree of Electronics Engineering.

Richard Demo Souza
Supervisor

Brazil

21/11/2018

The good thing about science is that it's true whether or not you believe in it - Neil deGrasse Tyson

Abstract

Keywords: Data Science, Student Performance, Education

This project aims to bring the use of data and factual analysis to the discussion about the current educational system's efficiency and effectiveness, aimed especially at Brazil's public superior education. It begins by first understanding what is already being done in other universities worldwide then applies those concepts to the current scenario at the Federal University of Santa Catarina. By using the concept of clustering, four main groups are created to identify the different profiles of students currently in the university. These clusters, or groups, help create an easy to understand framework of the initial steps that need to be taken to provide the necessary changes to improve results. It's easy to identify what areas need to be improved upon and what our current issues are.

List of Figures

List of Tables

Contents

Preface

University efficiency has been a matter of great concern especially for public-funded universities, which have the particular social responsibility of providing as many students as possible with equal opportunities. Typical lines of thought generally go in the direction of “Students need to perform better”, “Teachers are not providing the necessary help” or “University does not give the necessary support”. No matter which line resonates with largest number of people, one fact remains the same: these “conclusions” are mostly based on personal feelings or impressions. Even simple questions and metrics that should be of great concern for the university and for the general public, such as "How many students graduate? And in how long?", are often ignored and not even disclosed.

Is it possible to use very simple data models to have at least initial conclusions about the current scenario and provide some answers to the most common questions? This project aims to do just that, bring data to the discussions as means to provide a factual basis for which to discuss upon and provide decision makers with more information to take decisions on. By creating simple frameworks that facilitate the understanding of problems and opportunities, this project hopes to inspire and serve as an example of data analysis applied to education.

1 Introduction

1.1 Motivation

Brazil's education budget is far from being low. In fact, when compared with the other OECD members, Brazil has an above-average education expenditure in terms of GDP share[(3)]. Despite this, the actual results arising from such an important investment are scarce [(4)]. This is the outcome of an inefficient educational system that has many loopholes and lack of focus on results. In this context, studies that help understand the problems and limitations of Brazil's educational system become of great interest. Initial assumptions exist, mostly based on rumors and feelings, explaining the reasoning for this. But, in the field of data science it is often said that what cannot be measured, cannot be managed and a data driven approach to this problem has to exist, confirming or denying the rumors and establishing more confident responses.

To conclude, the main objective of this study is not to judge or have final conclusions on what should be done. Instead, the goal is to have a better understanding of the current scenario using data and creating simplified models that make it easier to understand. Thus, this study hopes to influence others in using a data driven approach to different kinds of problems faced not only in universities, but in the educational systems as a whole.

2 Literature Review

The use of Data Science in education has been growing steadily during the last couple of years with studies ranging from a more basic level, following up with retention and efficiency, to more complex applications such as the use of machine learning to create new possibilities for teachers and students[5]. In this chapter, a brief overview of these studies is presented.

2.1 The Study of Student Retention

Student retention is one of the main metrics reported in universities all around the globe. The reason for that is simple: it is highly correlated with efficiency and cost. One of the main examples of these types of studies is the "Dropout and Completion in Higher Education in Europe"[6], a study created by the European Union (EU) reporting the progress in three main metrics by all public universities in the EU:

- **Completion**: to have students successfully complete their study programme with a degree.
- **Time-to-degree**: to have students complete their study programme within a reasonable time period.
- **Retention or dropout**: the aim to have students re-enrol in a study programme until they complete their degree and to reduce the likelihood they drop out before completing their programme.

The objective of the report is aligned with Europe's 2020 strategy which is "to have at least 40% of 30-34–year olds complete higher education." In order to accomplish that, the number of students entering and leaving the system is of crucial importance.

In Brazil, the scenario is lagging behind when compared to most countries in Europe and North America. According to MEC[(7)], Brazil's Ministry of Education, the current target is to have 50% of young adults, with ages between 18 and 24 years old, enrolled in a higher level of education.

2.2 Applications of Data Science in Education

In general, applications of Data Science specifically for education have been shy. Education has often shown to be one of the last fields to adopt new technologies and innovations. This is evident by looking at the current formal educational systems worldwide, which have not had profound changes since the 19^{th} century with the Napoleonic Era [(8)]. While the reports mentioned in Section 2.1 show some progress in terms of using data for education, other more impactful changes are limited. Some other reports show that while collection of data has increased and is more accurate by the day, actual outcomes are still not present resulting in a lack of investment and confidence of what results can be brought from the use of data [(9)].

3 Methodology

This chapter briefly describes the tools and concepts used throughout this project, hoping to provide a simple basis as to understand the process followed to achieve the results.

3.1 Tools

Given the sheer amount of data that can be used in this project, a proper tool set has to be chosen. Thus, all of the analyses were carried out using Python[(10)], which is one of the most common languages used for works related to Data Science. Apart from Python, one other language that was used extensively was SQL (Structured Query Language)[(11)] to extract the data from the database. To join both of these tools and gather all information in one location, Jupyter Notebook[(12)] was used, which is an integrated development environment (IDE) that makes it easier to gather and share any of the results and data obtained throughout the analysis.

3.2 Sample Definitions

In Data Science, one of the common rules of thumb is to maximize the samples of data used. However, this proved to be a challenge in this project and some restrictions had to be made.

In order to improve quality of the data that is available, the analyses were all carried out for the course of Electrical Engineering at the Federal University of Santa Catarina instead

of Electronics Engineering. The reason for this is that the course of Electronics Engineering was started in August of 2009, meaning that there would be insufficient data to work with. The other restriction was that only data from students that enrolled after the first semester of 2000 was to be used. The idea was to reduce the number of curriculum changes, as there was a big one in 1999. There was also a limit on the enter date, which has to be before the first semester of 2010 to ensure sufficient time for the students to reach graduation.

3.3 Concept of Clustering

It is common knowledge that every student entering university has a different background, history, strengths and weaknesses, so of course many assumptions needed to be made to reduce the number of variables and make the results of this study easier to comprehend. To this end, a concept known as Clustering is utilized. Clustering is "the process of organizing objects into groups whose members are similar in some way"[(13)]. In other words, given only a part of the variables that describe the different students at the university, the question is how is it possible to find their similarities to further understand their behaviors and results during their years of studying?

In Figure 1 it is shown how the process of clustering works for a data set with only two variables. On the left side, all the points in the dataset are laid out with no specific pattern. On the right side some kind of pattern is found and highlighted by the different colors. The more variables are in, more robust and reliable the process tends to become.

To increase clustering reliability, one of the most important

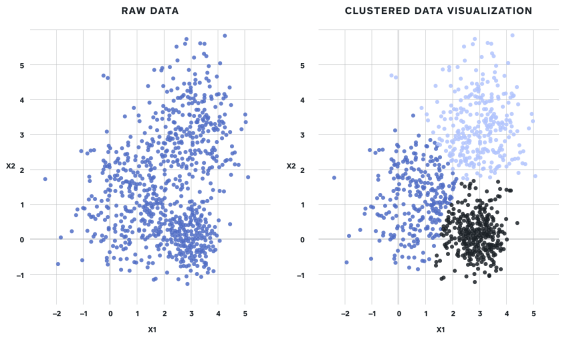

Figure 1 – Clustering Visualization[1]

steps is to find variables or features, that are good predictors of similarity. In other words, if the data set is analyzed variable by variable, is there a clear way to distinguish the different data points? This is commonly analyzed using the concept of correlation[2].

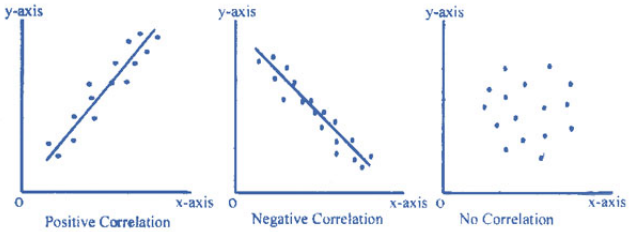

Figure 2 – Types of Correlation or Predictors of Performance[2]

The three types of correlation are shown in Figure 2. On the left-hand side, a positive correlation is illustrated, as the variable on the x-axis increases, the variable on the y-axis also increases. For a negative correlation the opposite occurs, as the variable on

the x-axis increases, the variable on the y-axis decreases. In the last case, at the right-hand side of Figure 2, a data set with no correlation is illustrated, meaning that the two variables are not good predictors of performance.

4 Results

In this chapter, some important definitions are made introducing the proposed concept of student success. Afterwards good predictors of success are indicated and finally applying the clustering concepts seen prior to form student groups.

4.1 The Definition of Success

As seen in Section 3.3, the first objective is to find variables that correlate or predict, student success. But before that, one very important question needed to be answered: what is student success? As previously seen in Section 2.1, many different metrics and definitions of success exist, however, this needed to be adapted to the reality at UFSC. Certainly, the definition of success is, in most cases,very subjective. For instance, a student could be considered successful if, after graduation, he/she finds a job. In the context of this work, the idea is to use something simpler as a definition of success, encompassing only the university.

At first, a definition was created based purely on experience and feeling in order to have some initial guidance on what should be pursued. The 1st definition included two terms:

1. **Graduation in under 6.5 years:** Initially it was believed that this was sufficient time to complete the course
2. **Maximum of three class failures:** This would include

most of the students and leave a considerable margin.

The first step after this initial definition was to get more insights into what the current scenario is. Table 1 shows the proportion of students that graduated, dropped out and still active among those considered for this study.

Metric	Description	Size	% Representation
Total Sample	Total number of students for the study	1077	100%
Graduated	Total number of students that graduated.	705	65.5%
Dropouts	Total number of students that dropped out	367	34.1%
Active	Total number of students that are still active in the course	5	0.4%

Table 1 – General context of current scenario

The next step was to extract time data: how long does it take for a student to graduate? What about to drop out? Table 2 shows the median and standard deviation of the time to graduate and time to dropout.

Metric	Description	Median	Standard Deviation
Time to Graduate	Time it takes for a student to graduate in semesters	11	2.5
Time to Dropout	Time it takes for a student to dropout in semesters	6	4.9

Table 2 – Time metrics of current scenario

From the results in Tables 1 and 2, an initial suspicion emerged: perhaps the criteria defined for success was not appropriate, because in general students did not take very long to graduate (when compared to the initial 13 semesters set) and in the end it seemed more a matter of graduating or not. This can be seen in Figure 3, where it is evident that the Time to Graduation follows a power law and not a normal distribution, ("The power law can be used to describe a phenomenon where a small number of items is clustered at the top of a distribution (or at the bottom), taking up 95% of the resources.")[(14)].

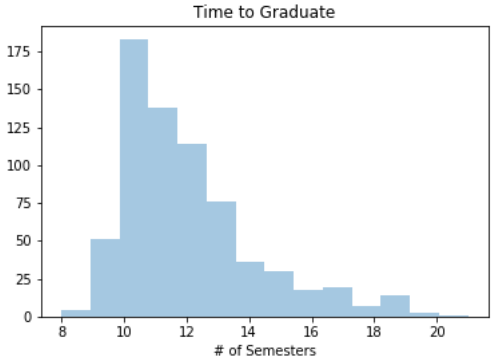

Figure 3 – Distribution of Time to Graduate

This distribution shows there is a high concentration of data very close to the median. Therefore it was decided that the **definition of success would be merely to graduate, not taking into account the time it took to graduate or the number of class failures.** This will become clearer after clustering the students into different groups. In fact the concept of success showed to be not very relevant.

4.2 Predictors of Performance

After closing the definition of success, there was the need to find general predictors of performance. This process begun by setting many initial hypothesis that would serve as guidance for the analyses. After a careful data analysis, the hypotheses that showed to be most relevant are:

1. Students that graduate have higher entry scores.

2. Students that graduate start better.

Numerous other hypotheses were created during the analyses, that showed to be not relevant for the prediction of this particular definition of success, as for instance:

- Students that graduate are younger;
- Students that graduate come from a specific quota;
- Students that graduate come from Santa Catarina;

4.2.1 Students that graduate have higher entry scores

To understand if students that performed better in entry exams had a higher chance of graduation we looked at three main scores: Math, Physics and Total scores. One thing that should be noted is that the database for exams scores is recent, therefore we do not have the data for all students. The Total is scored is composed of all the scores in the entry exam, ranging from 0 to 100, while the Math and Physics scores range from 0 to 10.

In Figures 4 and 5, one can notice a first distinction between students that graduated and students that did not. Students that graduated tend to have a higher score and are more concentrated on the higher scores (meaning a small variance) while students that did not graduate are more spread between all scores.

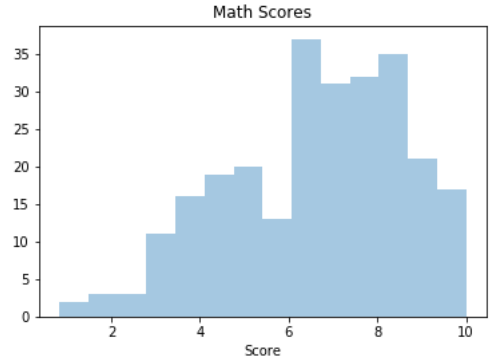

Figure 4 – Math scores of students that graduated

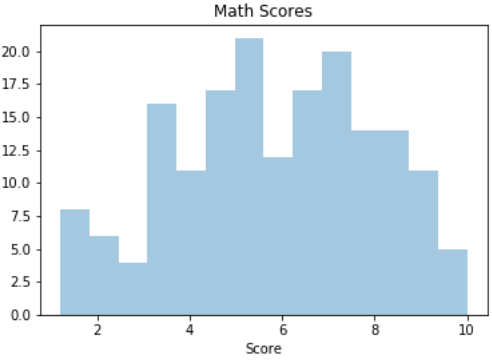

Figure 5 – Math scores of students that did not graduate

While students that graduated tend to have a median math score of 6.95, students that did not graduate have a median score of 5.92. This may not seem much at first, but by taking into account that the maximum score is 10, such an increase of 10% seems relevant.

This difference is even bigger when looking at Physics scores, the two histogram can be seen in Figures 6 and 7.

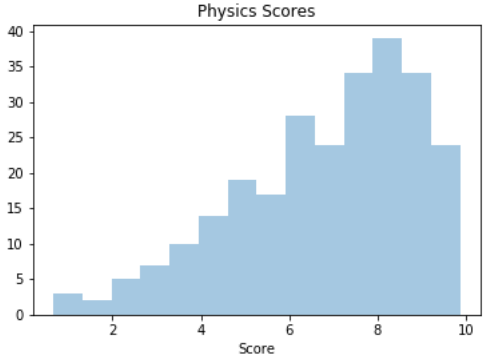

Figure 6 – Physics scores of students that graduated

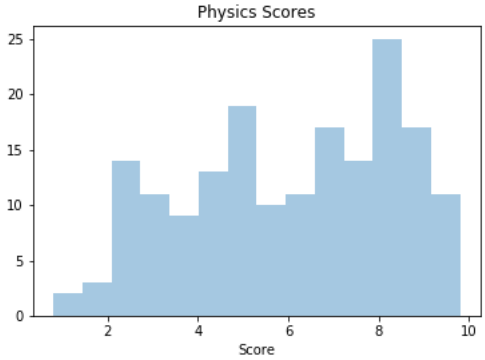

Figure 7 – Physics scores of students that did not graduate

Once again the difference between the physics score were of 1 point, 7.3 vs 6.3. However the same cannot be said about

the Total scores seen in Figures 8 and 9. The median difference between them is of only 0.8 points (66.0 vs 65.2) in a much larger point range (0 to 100). Table 3 shows the results in finer details.

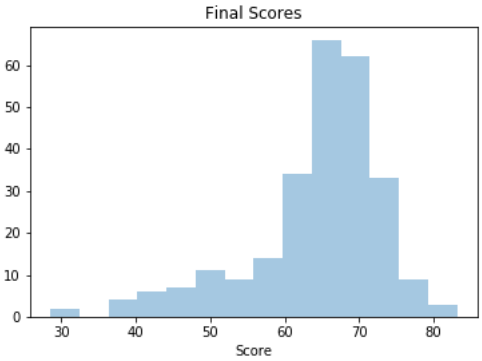

Figure 8 – Final scores of students that graduated

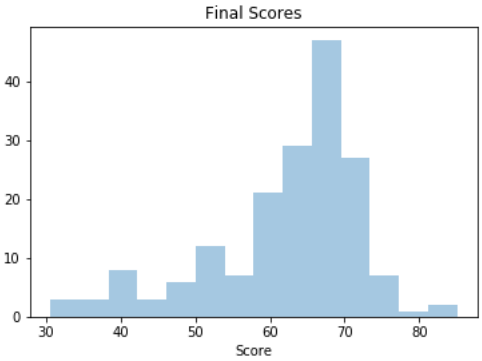

Figure 9 – Final scores of students that did not graduate

Type of Metric	Physics Score	Math Score	Final Score
Min	0.67	0.83	28.5
50 Percentile	6.8	6.6	65.96
75 Percentile	8.29	7.96	69.5

Table 3 – Scores for Entry Exam

Therefore our initial conclusion for this hypothesis is that, yes, Math and Physics scores are good indicators of performance while Final score is not so much. To further understand this situation an explanation for why some students have good scores but do not succeed needs to be formulated.

4.2.2 Students that graduate start better

The other main hypothesis created was that if the student showed good performance in the first semesters, there was a higher chance of graduation. In Figures 10 and 11, it can be seen a big difference between the first semester performance. Also the median student only fails 0.36 classes in the first semester. It's important to mention that the concept of failure is defined by the number of classes a student was enrolled in and the number of failures. Therefore is a student fails in 2 classes, the number is doubled.

Figure 10 – Failures in semester #1 of students that graduated

Figure 11 – Failures in semester #1 of students that did not graduate

This difference is so relevant that even though 20% of students fail more than 1 class in the first semester, only 28% of those eventually graduate. In fact if we were to expel all students that failed more than 1 class in the first semester, we would be

expelling 20% of all students but only 9% of those that graduate. This finding is one that shaped how the clusters were created.

To conclude this analysis, it was important to decide what definition should be used to describe a "good performance" in the first semester. As mentioned before, the average student that does graduate has a average failure of 0.36 classes in the first semester. With that in mind and the previous distributions, good performance in the first semester will be defined as 0 failures for the purpose of clustering.

4.3 Clustering Students

In Section 4.2, two of the best success predictors were pointed out: (1) Math and Physics scores in the entry exam and (2) first semester performance. However, one can notice that the first semester performance seems to be of greater importance. Therefore, four clusters were created based only on first semester performance and if the students graduated or not. Table 4 describes the four groups derived from the adopted clustering strategy as well as the number of students in each of these groups.

Clusters	**Description**	**Size (%)**
Group 1	Good first semester and graduates	535 (49.7%)
Group 2	Bad first semester and graduates	170 (15.8%)
Group 3	Good first semester and doesn't graduate	132 (12.2%)
Group 4	Bad first semester and doesn't graduate	240 (22.3%)

Table 4 – Cluster Definitions

In the next subsections each of the groups are broken up in further detail to understand behaviors and characteristics.

4.3.1 Group 1

As seen in Table 4, Group #1 is the biggest and it is also considered the best in terms of university efficiency. The two questions that need to be answered are: What is the pattern this group follows in university? Is it possible to understand the behavior prior to university with the limited available data?

First, it is possible to investigate the group's behavior considering Time to Graduation , Rate of Failures in all semesters, Number of Classes per semester, as shown in Figure 12, Table 5 and Table 6.

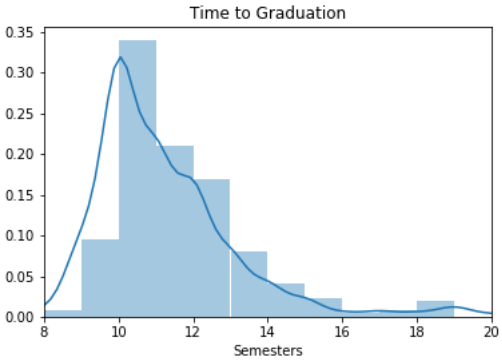

Figure 12 – Distribution of Time to Graduation of Group #1

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	0%	0%	0%	0%	0%	0%
75 Percentile	0%	0%	16.7%	20%	16.7%	0%

Table 5 – Rate of Failures during the first 6 semesters

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	6	5	5	5	5	5
75 Percentile	6	6	5	5	6	6

Table 6 – Number of Classes per semester of Group 1

The biggest insights that can be obtained from Figure 12 and Tables 5 and 6 is that students in this group graduate relatively fast (median value of 11 semesters), which is somewhat surprising considering the common belief inside the university. Moreover, it is possible to see that after Semesters 1 and 2, the failure rate rises a little and the number of classes tends to decrease as well (perhaps justifying the increased failures rate).

When observing their traits prior to the university, the only substantial difference found was in relation to their scores in the entry exams, seen in Table 7. Many other traits were also analyzed but nothing was found.

Type of Metric	Physics Score	Math Score	Final Score
Min	1.3	1.75	38.4
50 Percentile	7.52	7.25	67.2
75 Percentile	8.63	8.36	70.9

Table 7 – Entry exam scores of Group 1

When compared to the rest of scores (all except Group 1) there is a big difference: about 1 point in Physics and 1.5 in Math. Taking into consideration that the maximum score is 10, a 10% increase is very significant.

Type of Metric	Physics Score	Math Score	Final Score
Min	0.67	0.8	28.5
50 Percentile	6.4	5.94	64.5
75 Percentile	8.1	7.39	68.5

Table 8 – Score of all but Group 1

4.3.2 Group 2

Group 2 tends to have a bad first semester, which was seen as a predictor of success, but somehow still manages to graduate. What needs to be understood is how this happens and what is the group's behavior.

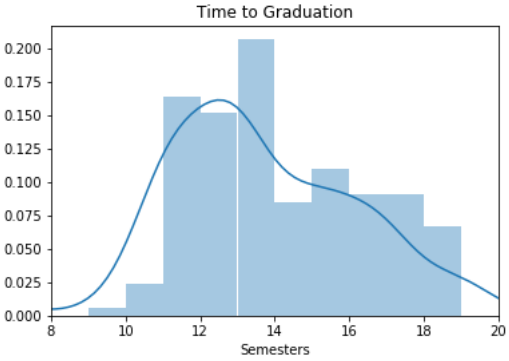

Figure 13 – Distribution of Time to Graduation of Group 2

As can be seen in Figure 13, the Time to Graduation of Group 2 tends to be much higher than Group 1, with median value of 13 semesters compared to 11 semesters for Group 1. The standard deviation is also much higher, which shows many students are graduating in more semesters than the median. The

reason for this becomes clear when looking at the failure rate and number of classes per semester, seen in Tables 9 and 10.

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	16.7%	20%	0%	20%	25%	20%
75 Percentile	33%	40%	33%	50%	44.6%	40%

Table 9 – Failure Rate for Group 2

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	6	5	3	4	4	4
75 Percentile	6	5	4	5	5	5

Table 10 – Number of classes per semester for Group 2

It is easy to see that Group 2 has a much higher failure rate throughout the whole time in university and at the same time a lower number of classes per semester. This explains the higher time to graduation.

Comparing the entry exam scores of Groups 1 and 2 it is possible to see a difference. It is considerably lower than Group 1 but very similar to the overall median.

Type of Metric	Physics Score	Math Score	Final Score
Min	0.67	0.83	28.5
50 Percentile	6.53	6.14	63.79
75 Percentile	7.79	7.15	67.0

Table 11 – Entry Exam scores for Group 2

4.3.3 Group 3

Group 3 has one the most curious patterns: a very good performance in Semester 1 but for some reason graduation is

not reached. The initial assumption for this group was that even though they had a good performance, they did not enjoy the course and decided to leave. To verify this there should be a small time to dropout. In other words, the students should dropout early, as for instance before Semester 4.

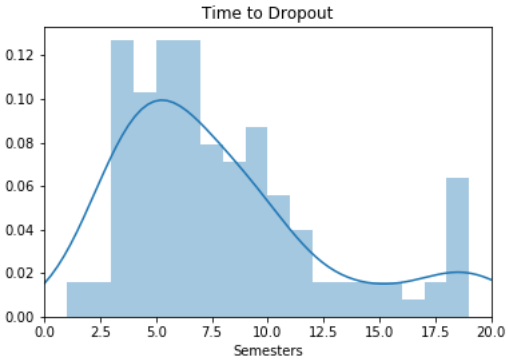

Figure 14 – Distribution of Time to Dropout of Group 3

From figure 14, one can notice that the initial hypothesis seems to be off. While the median time for the student to dropout is of 6.5 semesters (which is already very high when compared to our hypothesis) there are many other students dropping out after that. Therefore it was decided that two other subgroups should be created. Subgroup 3.1 includes students that dropped out before or on semester 6 while Subgroup 3.2 includes students that left after that. This should help with further understanding this group as a whole.

Type of Metric	Physics Score	Math Score	Final Score
Min	2.14	1.95	30.59
50 Percentile	6.6	6.63	67.62
75 Percentile	8.29	7.89	71.19

Table 12 – Entry Exam scores for Group 3

Scores for Group 3 are similar to the general scores, as seen in Table 12, but definitely not below the average.

4.3.3.1 Subgroup 3.1

Subgroup 3.1 includes 66 students which represents exactly 50% of Group 3. Because of their smaller time to dropout it was assumed that this subgroup includes students that did not enjoy the course and decided to leave.

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4
50 Percentile	0	0	40%	25%
75 Percentile	0	52.5%	75%	81%

Table 13 – Failure Rates for Subgroup 3.1

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4
50 Percentile	6	5	0	0
75 Percentile	6	6	4	2

Table 14 – Number of Classes per semester for Subgroup 3.1

The pattern that can be seen is that, even with good performances in Semesters 1 and 2, after a good start we see a decline of performance and lower number of classes leading to a dropout.

4.3.3.2 Subgroup 3.2

Subgroup 2 also contains 66 students but that take much longer to dropout.

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	0%	16.7%	8.3%	33.3%	50%	50%
75 Percentile	0%	33.3%	50%	66.7%	75%	75%

Table 15 – Failure rates for Subgroup 3.2

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	6	5	4	4	4	3
75 Percentile	6	6	5	5	5	5

Table 16 – Classes per semester for Subgroup 3.2

Subgroup 3.2 tends to have the behavior that in general is not desired for the university when thinking in terms of efficiency, which is due to the long time to dropout. It is also easy to notice in Table 16 that their performance tends to get worse over time.

4.3.4 Group 4

Even though Group 4 has a worse performance in Semester 1, it also has a similar time to dropout in relation to Group 3. The difference is the median time to dropout of 5 semesters instead of 6.5. Therefore, Group 4 was also divided into two subgroups. Subgroup 4.1 contains students that left before or on Semester 5 while Subgroup 4.2 contains students that left after that.

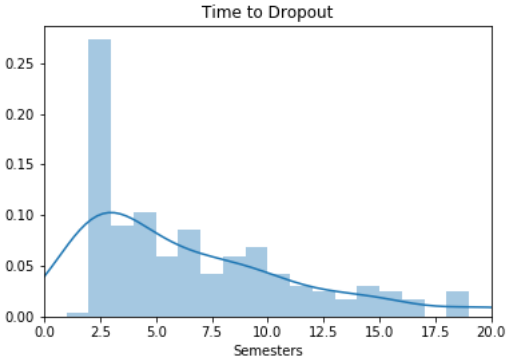

Figure 15 – Distribution of Time to Dropout of Group 4

Type of Metric	Physics Score	Math Score	Final Score
Min	0.8	1.2	31.29
50 Percentile	5.56	5.55	63.4
75 Percentile	8.1	7.2	67.49

Table 17 – Entry Exam scores for Group 4

Group 4 tends to have a bit lower exam scores but nothing much different to other groups.

4.3.4.1 Subgroup 4.1

Subgroup 4.1 contains a total of 124 students which represents 51.6% of Group 4. What we see in Subgroup 4.1 is that basically most of the students leave in Semester 1 or 2.

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4
50 Percentile	6	0	0	0
75 Percentile	6	3	0	0

Table 18 – Classes per semester for Subgroup 4.1

At first it was not known why they leave so quickly, but when looking at the classes they were failing, it became clear that their bad first semester performance was not due to difficulties with the classes and was just actually they, leaving for something else. They have similar failure rates for diverse classes, such as Calculus I and Technical Drawing, whereas other groups have very distinct failure rates for these disciplines.

4.3.4.2 Subgroup 4.2

Subgroup 4.2 contains a total of 116 students, which represents 48.3% of Group 4. The pattern that was imagined for this specific group is of students who struggle with initial performance and eventually decide to leave.

Type of Metric	Semester #1	Semester #2	Semester #3	Semester #4	Semester #5	Semester #6
50 Percentile	33.3%	33.3%	50%	50%	66.7%	66.7%
75 Percentile	33.3%	60%	75%	75%	100%	100%

Table 19 – Failure rates for Subgroup 4.2

This is indeed the pattern that is observed: an increase in failure rates and lower number of classes per semester, leading to an eventual dropout.

5 Concluding Remarks

This section futher analyses the results laid out in the previous chapter.

5.1 Analyzing the Results

Figure 16 summarizes all groups seen in other sections. After the clustering of these groups and obtaining some familiarity with them, it is fundamental to reflect on what they really mean in terms of impact on the university and give hints on what actions should be taken in order to improve retention rates.

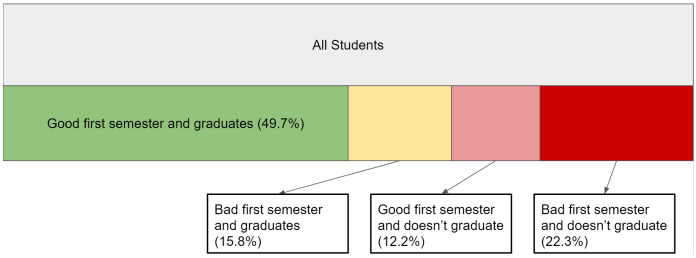

Figure 16 – Summary of all groups

Let us start with the easiest group to understand, Group 1. They show good performance in Semester 1, very small failures rates in subsequent semesters and eventually graduate in a good time frame. After Semester 2, they do have localized higher rates of failures but still much lower than any other group. An interesting finding by exploring this group is that it defies what

is commonly believed in the university, that students take a long time to graduate and are generally not very efficient. However what can be seen here is that most of the students that graduate, do it in a generally acceptable number of semesters. Further exploration of this group could try to find some other patterns that could aid in the early identification of it.

Group 2 becomes a bit harder to analyze. It was seen in Section 4.3.2 that they do not have a good first semester but eventually graduate, however in a much longer time frame. While they do have lower entry exam scores, it would be difficult to confirm that this is the only reason for their behavior. Questions such as "Why do students with high scores sometimes fall in this group?" are not easy to be answered. An attempt to understand this would be to see in which classes they are failing. Table 20 shows the classes with lowest approval rates for the students in Group 2. The rates of these classes are similar to those observed for other groups that are failing, but still important to be known.

Class	Approval Rates
Electrical Circuits	50%
Calculus I	52%
Linear Algebra	53.5%
Microprocessors	55.5%
Scientific Computing II	58%

Table 20 – Approval rates for Group 2

Group 3 is even more interesting. From a quick analysis of first semester performance and exam scores, it seems that they will graduate, but however they do not. By breaking this Group into two it became easier to understand the two very

different student profiles. In Subgroup 3.1, students are leaving very early in the course and after Semester 2 and 3 they are basically out. A qualitative analysis on top of this data could indicate that Subgroup 3.1 is probably composed of students that did not like the course and decided to leave even when performing well. Subgroup 3.2, however, is more difficult to be understood. Students in this group start with good performance but with time it declines and only after Semester 6 they start to dropout. For now that is all the information that is available, even when analyzing the classes they fail, there is no clear pattern.

Group 4 represents a total of 22.3% of all students, even though common knowledge at the university would often bring this number much higher. At first it was thought that this group would be composed of students that did not perform well and then eventually just gave up. But after breaking it into two subgroups, something else was found. Subgroup 4.1 basically leaves in the first semester and they have failures rates that are equal or similar for all classes. This indicates that they are just leaving and not actually failing the classes. Subgroup 4.2 is very similar to Subgroup 3.2: the performance worsens over time and eventually they leave.

To summarize, if actions were to be taken based on this analysis the most important points to take away would be:

- How can we further understand Group 1 as to improve the university entry process as well as the course as a whole in order to increase its proportion?

- Is there any way to give stronger support for students in Group 2, allowing them to progress faster and graduate in

less time?

- Subgroup 3.1 has a good initial performance but leaves, what can be done to prevent and further understand this?
- Subgroups 3.2 and 4.2 take a very long time to dropout, why is this? Is there any way to either support them or help them decide faster? Should students be allowed to stay so long with poor performance given that only a small amount of students with bad performance graduate?
- Given that first semester performance has shown to be so relevant, is there any way to take actions after the first semester?

5.2 Final Remarks

It has been mentioned many times throughout this document that retention and approval rates need to rise but it is crucial to remember that what cannot be measured, cannot be managed. This project had the objective of using data to get at least a little bit further understanding for the current scenario at Electrical Engineering at the Federal University of Santa Catarina. This should help provide some initial insights, by clustering and grouping all the students in the university. The simplifications and clustering made in this analysis provide an easier approach to understanding the problems. Hopefully this serves as an incentive for further actions and improvements in the university.

Bibliography

1 KALER, I. *So You Have Some Clusters, Now What? – Square Corner Blog – Medium*. <https://medium.com/square-corner-blog/so-you-have-some-clusters-now-what-abfd297a575b>. (Accessed on 11/02/2018).

2 SURVEY SYSTEM. *Correlation - Statistical Techniques, Rating Scales, Correlation Coefficients, and More - Creative Research Systems*. <https://www.surveysystem.com/correlation.htm>. (Accessed on 11/02/2018).

3 OLIVEIRA, E. *Globo - Brasil gasta mais em educação em relação ao PIB que a média de países desenvolvidos*. <https://oglobo.globo.com/sociedade/educacao/\brasil-gasta-mais-em-educacao-em-relacao-ao-pib-que-media-de-paises-desenvolvidos-22858629>. (Accessed on 11/11/2018).

4 FAJARDO, V. *MEC - 7 de cada 10 alunos do ensino médio têm nível insuficiente em português e matemática*. <https://g1.globo.com/educacao/noticia/2018/08/30/7-de-cada-10-alunos-do-ensino-medio-tem-nivel-insuficiente-em-portugues-e-matematica-diz-mec.ghtml>. (Accessed on 11/11/2018).

5 DHAR, V. *Data Science and Prediction*. <https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/abstract>. (Accessed on 11/22/2018).

6 EUROPEAN UNION. *Dropout and Completion in Higher Education in Europe*. 2015. <https://supporthere.org/sites/default/files/dropout-completion-he_en.pdf>. Accessed on 2018-10-02.

7 MINISTRY OF EDUCATION. *Planning the Next Decade*. 2014. <http://pne.mec.gov.br/images/pdf/pne_conhecendo_20_metas.pdf>. Accessed on 2018-10-02.

8 NAPOLEON SERIES. *The Revolution, Napoleon, and Education.* <https://www.napoleon-series.org/research/society/c_education.html>. (Accessed on 11/11/2018).

9 STONE, A. *Will Data Scientists Have a Big Impact on Education?* <http://www.govtech.com/education/k-12/Will-Data-Scientists-Have-a-Big-Impact-on-Education.html>. (Accessed on 11/11/2018).

10 PYTHON. *Welcome to Python.org.* <https://www.python.org/>. (Accessed on 11/02/2018).

11 QUINSTREET. *SQLCourse - Lesson 1: What is SQL?* <http://www.sqlcourse.com/intro.html>. (Accessed on 11/11/2018).

12 PROJECT JUPYTER. *Project Jupyter.* <http://jupyter.org/>. (Accessed on 11/11/2018).

13 DEIB-POLIMI. *Clustering - Introduction.* <https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/>. (Accessed on 11/02/2018).

14 DATA SCIENCE CENTRAL. *Power Law and Power Law Distribution - Statistics How To.* <https://www.statisticshowto.datasciencecentral.com/power-law/>. (Accessed on 11/02/2018).